ATLANTIS PRESS

*Advances in Intelligent Systems Research (AISR), volume 155*

2018 International Conference on Advanced Control, Automation and Artificial Intelligence (ACAAI 2018)

# Website Intelligent Recommendation Based on K-means and Apriori Algorithms

Shaohua Zhang, Changhua Liu* and Qiaodan Li

School of Mathematics & Computer Science, Wuhan Polytechnic University, Wuhan, Hubei, China

*Corresponding author

***Abstract*—The recommended algorithm is one of the most popular applications of today. Firstly, the original data is cleaned and processed, and then the association rules model and user value analysis model are established in this paper. Secondly, a Apriori algorithm is used to analyze the relationship between user history access records and the user group of K-means algorithm is used to divide value. Finally, the experimental results show that the results of the output of the association rules and the clustering analysis of the user value have some reference significance.**

***Keywords − Recommendation; Association rules; Apriori; K-means***

## I. INTRODUCTION

With the rapid development of Internet technology at home and abroad, the research and application of the proposed algorithm are changing rapidly. For association rules mining is an important research direction of the academic circles. Among them, Wal-Mart's beer and diaper merchandise collocation strategy is the classic commercial case that used the methods of data analysis and association rules to create tremendous commercial value. Shenyi Qian[1] proposed a four-layer mining model, combined with K-means algorithm and Apriori algorithm, to construct a new feature word extraction method to classify the scientific literature.The method reduces the information loss in the clustering process, and can find the feature words more accurately in the document corpus. Meiling Liu[2] put forward a kind of clustering algorithm based on maximum frequent itemsets,

combining correlation analysis and cluster analysis, make full use of the data items in the cluster, has higher clustering accuracy and efficiency of algorithm. Yue He's[3] popularity index model is established for sina weibo data,

classifying all users using K-means algorithm and then use the Apriori algorithm analysis topic category, the correlation between characteristics of hot topic in different categories of users and user's relevance between different categories topic has achieved good result.

In general, research on association rules and clustering algorithms is of great research significance in the field of data mining[5]. According to the data mining of the National Undergraduate Data Mining Contest website provided by TipDM Company, this paper analyzes the relationship between the historical user access records through the establishment of the association rules model and the user value analysis model. Not only can the relationship between historical user access records be analyzed, but also the user group value is classified in historical data.

## II. BUILD THE MODEL

### C *Apriori*

In the field of association rules algorithm, the most classical and basic is Apriori algorithm, which use layer by layer search and iterate computation[6,7]. The nature of the Apriori algorithm is that if a term is frequent in the item set, all of its subsets are frequent[8]. I={ $i_1, i_2, \cdots, i_m$ } is a collection of all items in D. D represents the total number of tasks in all the database transaction sets, and the transaction T represents the total number of tasks D set. Let X be the set of items in I. If $X \subseteq T$ , then call

the transaction T contains the item set X. Now X and Y represent two concrete item sets, and $X \subset I, Y \subset I$ , $X \cap Y = \Phi$ and $X \Rightarrow Y$ is called an association rule that satisfies certain relations. Support is used in the transaction database item set in all transactions X and Y simultaneously appear in the probability of memory[9]:

$$S(X \Rightarrow Y) = P(X \cup Y) = |\{T : X \cup Y \subseteq T, T \subseteq D\}| / (D(X)) \times 100\% \quad (1)$$

Confidence level refers to the total probability of all occurrences of Y in all itemsets in the transactional database as:

$$S(X \Rightarrow Y) = P(X \cup Y) = |\{T : X \cup Y \subseteq T, T \subseteq D\}| / (S(X)) \times 100\% \quad (2)$$

### B *Apriori Algorithm Execution Flow Chart*

In this paper, the input parameters of the model are selected as the minimum support of 1% and the minimum confidence of 72%.

Copyright © 2018, the Authors. Published by Atlantis Press.



This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

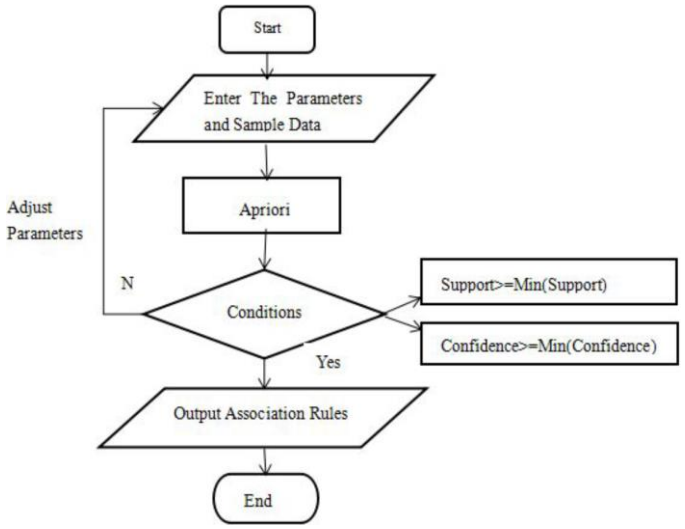


FIGURE I. APRIORI FLOW CHART

### C *K-means*

The k-means algorithm is also known as K mean, which is a clustering algorithm based on division[8]. By specifying specific K values in advance, this algorithm divides the data sets that need clustering into K different clusters. K-means algorithm execution steps[10]:

1)The initial clustering center is chosen arbitrarily from D. The center is any k objects.

2)Iteration calculation.

3)Calculate the center object of clustering, that is, the mean value of clustering objects, compare the distance between each object and the center object, and then divide the corresponding objectsagain, and the division is based on the minimum distance in the alignment.

4)Find a cluster center object again.

5)Until you find no new center object.

## III. EXPERIMENT

According to the characteristics of datasets and related business knowledge for users to access data based, we gather it into three categories: enterprise pages, enterprise application pages and web data mining competitions. Finally, the user group are divided into group A, group B and group C.

### A *Experiment Environment*

In this paper, Matlab R2014b is useded to realize programming.Experiment environment:

Processor:Inter (R) Core (TM)

CPU: i3-2330M

Memory: 8G

Frequency: 2.20GHz

Operating System: Windows 7-64 bit

### B. *Experimental Data Preprocessing*

In this paper, a half-year time period is selected as the observation window, and all detailed records of the users in the window are extracted to form a required data sample set. The

data includes attributes such as user ID, access IP, access time, access page, keywords, primary and

secondary tags, source website, origin website, sessionID. Some of the original data is shown in table 1.But access to source IP, access time, keywords, web sites and source web pages has nothing to do, so we need to drop these attributes. The data after the attribute specification is shown in table 2. Data mining and training net pages number attribute are the number of web pages for users to visit secondary label case tutorial, teaching resources and training information. The enterprise applies the web pages for users to access the level 2 tag innovation technology and enterprise applications. The data mining contest web page number is the number of pages for users to visit the first level match and the evaluation and

TABLE I. DATA PREPROCESSING

| User Id | Session Id | Ip | Time | Page | Keyword | Label(level 1) |
|---|---|---|---|---|---|---|
| 200083 | 5bceca6 -5a39-4jd | 113.96.8.218 | 1/29/2016 17:37:09 | http://www.tipdm.org/ | Practical Case Studies | Educational Resources |
| 200085 | fd382d7 -a08f-jf5 | 203.43.116.99 | 1/29/2016 18:37:17 | http://www.tipdm.org/ qk/564.html | Data Mining Title | Competition |

TABLE II. DATA ATTRIBUTE SPECIFICATION

| User Id | Session Id | Page | Keyword | Label(lavel 1) | Label(level 2) |
|---|---|---|---|---|---|
| 200083 | 5bcec0a6-5a39-4jd | http://www.tipdm.org/ | Practical Case Studies | Educational Resources | Modeling Tools |
| 200085 | fd382d7-a08f-jf5 | http://www.tipdm.org/index.html | Data Mining Title | Competition | Case Tutorials |

TABLE III. ASSOCIATION RULE RESULTS

| Id | Front | Front key words | Back | Front key words | Support | Confidence |
|---|---|---|---|---|---|---|
| 1 | http://www.tipdm.org/ts/535.html | Mathematical Modeling | http://www.tipdm.org/sj /560.html | Data Analysis | 3.43% | 78.32% |
| 2 | http://www.tipdm.org/sj/560.html http://www.tipdm.org/sj/535.html | Data Analysis Data Mining Study | http://www.tipdm.org/sj /556.html | Mining Tools Tutorial | 1.75% | 74.32% |

### C. *Association Rules Model Analysis*

Using Apriori algorithm to mine user's access data, some results are shown in table 3. For mining association rules in the serial number 2, the model of the output support at 1.75%, indicating the user page in access to data analysis and data

mining under the premise of case analysis page to visit mining tools use a 74.32% chance of this tutorial.

### *D. User Value Analysis*

The discrete standardized processing of selected cluster centers is shown in table 4. The results are as follws.

TABLE IV DISCRETIZATION PROCESSING OF USER VALUE MODEL

| Category Cluster | Cluster Number | Cluster center | | |
|---|---|---|---|---|
| | | Training Pages | Contest Pages | Enterprise Application |
| A | 28 | 0.80673 | -0.2632 | 0.59726 |
| B | 11 | 0.42375 | 2.1981 | 0.09097 |
| C | 46 | -0.5923 | -0.3654 | -0.3853 |

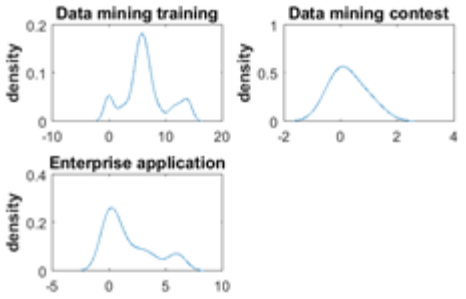


FIGURE II EXPERIMENTAL RESULTS

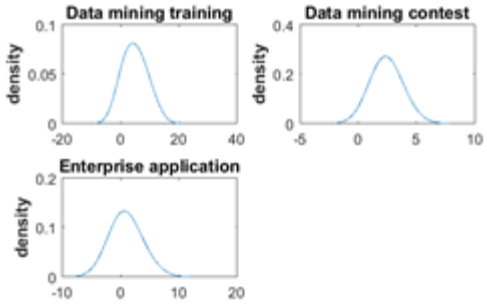


FIGURE III EXPERIMENTAL RESULTS

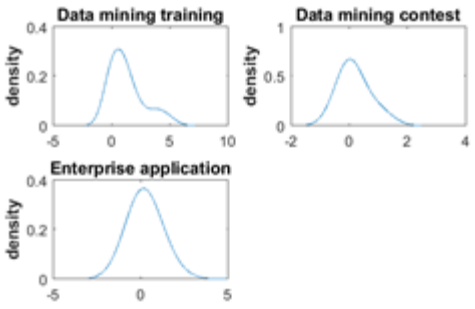


FIGURE IV. EXPERIMENTAL RESULTS

From the experimental results, for Figure 2, the number of data mining training pages in user group A are between 0~15, the number of data mining competition pages are between 0~2, and the number of web pages applied by enterprises are between 0~6. In Figure 3, the data mining training web pages applied to group B's visits are between 0~20. The data mining competition pages are between 0~7, and the web pages applied by enterprises are between 0~10. In Figure 4, the data mining training web pages applied to user group C's visits are between 0~6. The data mining competition pages are between 0~3, and the web pages applied by enterprises is between 0~4. For user group A, its data mining training and competition webpage access can know that it is quite interested in data mining, We can recommend corresponding training courses to users. For the user group B, we can know that the group is similar to an enterprise users, and we can recommend an enterprise application for data mining to them. For user group C, we can know that the user group is not interested in data mining, so we can recommend some data mining cases to them, and cultivate their interest in data mining.

## IV. CONCLUSION

Based on the data mining analysis of TipDM data mining competition website access, this paper makes an in-depth research on the application area of data mining. First, the association rule algorithm and clustering algorithm are used to analyze the data, and the association rules model and the user value model are constructed. Secondly, the association rules are used to analyze the users' access records, and the degree of association between the access pages is excavated. At last, the cluster analysis model is used to analyze the user's group label, and the related service is recommended to the user groups with different labels.

## REFERENCES

[1] Shenyi Qian, Yanling Zhu, Haodong Zhu,"Multi-level feature extraction method based on K-means and Apriori", Journal Of Central China Normal University(Nat. Sci), 2015, 49(3): 357-362.

[2] Meiling Liu, "Clustering Algorithm Based on Maximal Frequent Itemsets", Computer Engineering, 35(17), 2009, 43-45.

[3] Yue He, Yue Zhang,"On the User Characteristics of Different Topics on Sina Microblog", Journal of Intelligence, 2016, 35(7): 107-110.

[4] Shouning Qu, Caiyun Dong, Dejun Xu, Tong Wu, "Research on Algorithm of Association Rules and its application in Education System", Computer Systems & Applications,2005,4: 20-23.

[5] Xinhua Huang, "Design and Implementation of A Book Recommendation System based on Apriori and K-means algorithms", Hunan University, 2016.

[6] Yan Shen,"The Research of High Efficient Data Mining Algorithms for Massive Data Sets", Jiangsu University, 2013.

[7] Zhongtao Jia, "Research and Implementation of Personalized Movie Recommendation System", Southwest University, 2015.

[8] Jiangyue Liu,"Research and implementation of teaching analysis system based on K-means and Apriori algorithms", Nankai University, 2011.

[9] Agrawal R, Srikant R. "Fast algorithms for mining association rules in large databases"//Proceedings of the 20th International Conference on Very Large Databases, Santiago, Chile, 1994, 487-499.

[10] Najadat H M, Al-Maolegi M, Arkok B. "An improved Apriori algorithm for association rules". International Research Journal Of Computer Science and Application, 2013, 1(1): 1-8.